Finding Predictive Relationships between Notifiable Diseases with Markov Blanket Discovery

ERNEST MWEBAZE[3]
JOHN A.QUINN
Faculty of Computing and IT, Makerere University

Abstract

Most countries collect surveillance data at least weekly on several notifiable diseases. This data is to inform policy formulation as well as aid in forward planning. Knowing the relationships between these diseases is not only helpful for preventative planning but for predictive purposes as well. In this paper we present a computational method for analysing the relationships that exist amongst these diseases. We employ the Incremental Association-Based Markov Blanket algorithm to find sets of mutually informative diseases with high predictive power in a dataset covering the 80 districts of Uganda. Results indicate sufficiency of our methods in obtaining non-spurious relationships amongst notifiable diseases. The predictive relationships we identify could be used to improve the accuracy of existing disease monitoring systems.
Categories and Subject Descriptors: I.2.1 [**Computing Methodologies**]: Artificial Intelligence– Medicine and Science.

General Terms: Disease Association learning, Causal Analysis, Markov Blanket Analysis, Notifiable diseases, Biosurveillance.

IJCIR Reference Format:

Ernest Mwebaze and John A. Quinn. Finding Predictive Relationships between Notifiable Diseases with Markov Blanket Discovery. International Journal of Computing and ICT Research, Special Issue Vol. 4, No. 1, pp. 30 - 36. http://www.ijcir.org/Special-Issuevolume4-number1/article4.pdf.

1. INTRODUCTION

Work on monitoring outbreaks of disease has recently focused on using alternative sources of data, such as in symptomatic monitoring (e.g. looking at sales of over-the-counter drugs, absenteeism and hospital admissions for related illnesses). Working out how to take advantage of these extra sources of information makes the task of inferring the extent of disease more feasible when direct measurements may be scarce and untimely. In this work, we consider the information about disease rates that can be extracted from data about the prevalence of *other* diseases. Because some groups of diseases share common factors, there can be predictive relationships between disease counts. This extra source of information has the potential to improve the accuracy of existing disease monitoring systems in situations where other data is limited.

An analysis based only on direct correlation is not adequate for this task, as this is prone to finding spurious associations. Consider for example two random variables we might measure in Uganda: 'death by plague' (which has been steadily decreasing), and 'mobile phone access' (which has been steadily increasing). The two are strongly (anti-)correlated, but each has no predictive power with respect to each other. However, using more sophisticated techniques and with some conditions on the covariates we have available, we can find the smallest set of covariates with the greatest predictive power – without being misled by such spurious associations.

[3] Author's Address: Ernest Mwebaze and John A. Quinn, Faculty of Computing and Informatics Technology , Makerere University, Kampala, Uganda, [emwebaze, jquinn]@cit.mak.ac.ug

International Journal of Computing and ICT Research, ISSN 1818-1139 (Print), ISSN 1996-1065 (Online), Special Issue, Vol.4, No.1, pp. 30 - 36, October 2010.

In probabilistic graphical models, the minimal set of variables which jointly determine the distribution of a variable is known as the Markov blanket of that variable.

We employ the Incremental Association-Based Markov Blanket (IAMB) algorithm [Tsamardinos and Aliferis 2003] which is applied in order to discover Markov Blankets (MB) for each disease. MB discovery techniques in general try to find the set of features that constitute the immediate family of a particular feature and these include the parents of the feature (causes) the children (effects), and the variables sharing a child with that feature. This set (the MB) contains all the features that have an effect on that particular feature due to the associative nature of their linkages.

For epidemiological surveillance data it is important to know what associative relationships exist between the diseases as this aids prediction, planning and management. It is worth noting that these associative links may exist because of extra latent or hidden variables that may be causing both diseases in an associative relationship. This is common in disease data because like in our case, only information about the diseases is collected, other relevant data that may constitute the latent variables e.g. climatic conditions, sanitation conditions, movement of people etc are not collected yet with disease causation and transmission these are crucial.

The Uganda Health Management Information System (HMIS) is the principle tool used by the Ministry of Health in Uganda to collect all health related data in the country for purposes of planning, managing and evaluating the health care system in Uganda. The HMIS specifies specific periodic data collection routines for all health centers in the country to collect data on specific diseases and other health related items. Most of the analysis of this data is done using basic descriptive statistics. In this paper we present some computational techniques from recent research in computing and try to address them to this data for purposes of a more general associative analysis.

In the next section we discuss the data we used and the notifiable diseases we focused on. We further discuss the associative techniques in the later sections and then give results.

2. DATA

A notifiable disease is any disease in a particular country whose every incidence has to be reported to relevant government authorities. The aim of this is usually to enable the government monitor the disease effectively as well as strategize on how to handle the disease especially if an epidemic is suspected.

The HMIS has a form that records weekly the number of new cases of a particular notifiable disease and the number of deaths from a particular disease for the previous week. The data used in this paper comprises over seven years of this data from 80 districts of Uganda and has just over 30,000 records. Preprocessing steps for the data involved removing all records with missing values. For this study we also considered only the weekly disease incident counts (not the deaths) of the following diseases in the HMIS 033b; Acute flaccid paralysis (Afpn), Rabies (Rbn), Cholera (Chon), Dysentry (Dysn), Guinea worm (Gwmn), Malaria (Maln), Measles (Mssn), Meningococcal Meningitis (Menn), Neonatal tetanus (Nntn), Plague (Plgn), Ty-phoid (Typn) and Sleeping Sickness (Slsn).

3. CAUSAL ASSOCIATIVE ANALYSIS

Causal associative analysis deals with finding associations amongst the diseases that are causally oriented. In recent years causal structure discovery has emerged as a branch of machine learning. The goal is to learn causal relationships from purely observational data, without being able to perform manipulations. The first benefit of such techniques is the ability they provide to understand causal mechanisms in domains such as food security, epidemiology or genetics where it is either expensive, unethical or impossible to carry out certain experiments. Other advantages of causal analysis include (i) prediction under interventions: by knowing which variables are either causes or effects of a target variable, we can make robust predictions on future data even when interventions are made on some of the variables, (ii) manipulation: predicting the consequences of given interventions due to an external agent on the natural system, and (iii) counterfactual: given a certain outcome was observed, predicting what would have happened if different action had been taken.

It is critical to note that causal analysis is of an entirely different nature to correlation-based analysis (density modeling) of a dataset, as covered in [Pearl 2000]. A causal relationship implies a change in one variable will have a corresponding change in any variable that is causally dependent on it, and this

must hold even when the variables are subjected to external interventions. This is very critical for example in prediction where the data is subject to external interventions. For example, if we consider the correlation between time spent in bed and the death rate. Studying these two variables would show a positive correlation between them. It would hence seem like the most obvious conclusion is spending time in bed 'causes' death and the converse not sleeping at all ensures you live forever. However if we apply an intervention to the system, e.g. forcing people to spend more time in bed, there probably would be no increase in the death rate. A close analysis would indicate that there is another variable, sickness for example, that causes both an increase in time spent in bed and the death rate.

In this study we are however not interested in the direction of causation of the relationships between the different diseases. We are more interested in whether there happens to be a link between any two diseases that is causally motivated. This is because with disease data of this nature there are bound to be a potentially high number of latent variables that causally affect any of the diseases or groups of diseases hence determination of the direction of causation becomes more difficult. For this study also, just knowing there exists some form of relationship between any two diseases is of critical importance as earlier highlighted.

3.1 IAMB

The Incremental Associative Markov Blanket method is a constraint-based causal analysis method that uses conditional independence tests to determine the relationship between variables, and so determine the Markov Blanket. Formally the Markov blanket of a target variable T, MB (T), is the minimal set of variables or features such that every other variable is independent of T given MB (T) [Tsamardinos and Aliferis 2003].

The IAMB algorithm is an improved version of the Grow-Shrink (GS) algorithm [Margaritis and Thrun 1999]. Essentially these algorithms are two phased. In the first phase the algorithm finds a set of all relevant variables to the target and then in the second phase conditional independence tests are used to trim down the set of relevant variables to the Markov blanket. The GS algorithm has some limitations for example it requires a sample at least exponential to the size of the Markov Blanket and cannot scale to thousands of variables [Guyon et al. 2007]. IAMB has been proven correct and sound for discovering Markov Blankets [Tsamardinos et al. 2003] and has since produced several variants of itself including Fast-IAMB [Yaramakala and Margaritis 2005], IAMBnPC, parallel IAMB, etc.

Given a dataset D, the IAMB algorithm will discover a unique Markov Blanket MB (T) if (i) all the data, D is generated by process that can be faithfully represented by Bayesian Networks, and (ii) if there exist reliable, statistical, conditional, independence tests and measures of association for checking independence and strength of association of T with some other variable X given a set of variables

Y. If (i) and (ii) do not hold then the output obtained is a heuristic approximation of the Markov Blanket of the target T.

Algorithm 1 IAMB Algorithm

Input:
D - Dataset
T - Target variable

Assign $\mathbf{MB} = \emptyset$

Phase I :
Repeat
  Find X, $argmax_{x \in (X - \{T\} - \mathbf{MB})}$ mutualInformation$(X, T \mid \mathbf{MB})$
    IF NOT condIndependent$(X, T \mid \mathbf{MB})$
    Add X to **MB**
Until MB does not change.

Phase II :
For $X \in \mathbf{MB}\,(T)$ do
  IF condIndependent$(X, T \mid \mathbf{MB} - \{X\})$
    Remove X from **MB**
Return MB

The IAMB algorithm is shown in Algorithm 1. It starts with an empty set for the MB. In phase 1 (also called forward or grow phase) it incrementally admits members, X into the set MB if X has the largest association with T conditioned on the MB. In phase 2(backward or shrink), conditional independence tests are used to trim all the false positives from the set MB, obtaining the real Markov Blanket.

3.2 Significance Testing

Since our data is of a continuous nature, we used Fisher's Z-test as a measure of conditional independence with varying significance levels α. By varying the significance levels of the tests we were able obtain different MBs corresponding to different strengths of association between the different diseases.

4. RESULTS AND DISCUSSION

Causal Association learning was carried out for the 12 diseases/features in the dataset for various significance levels. Figure 1 depicts a graph showing the associations amongst the various diseases for the default significance level of 0.05. As can probably be expected, the graph depicts a highly connected graph with multiple relationships between the diseases. The links between the diseases represent associative relationships obtained when thresholding the selected Markov Blankets per disease at a significance level of 0.05.

Fig. 1: Association graph showing the associative linkages between the notifiable diseases represented as : Dysn -Dysentery, Typn -Typhoid, Gwmn -Guinea Worm, Nntn -Neonatal Tetanus, Menn -Meningococcal Meningitis, Plgn -Plague, Mssn -Measles, Afpn -Acute Flaccid Paralysis, Chon -Cholera, Slsn -Sleeping Sickness, Rabn -Rabies and Maln -Malaria.

Figure 2 shows the actual Markov Blankets obtained for each of the 12 notifiable diseases at a significance level of $\alpha = 0.05$. The order of the diseases in the Markov Blanket in a sense represents the relative strength of association with the target disease. Most of the linkages make heuristic sense for example there appears to be a link between diseases related to sanitation for example dysentery, typhoid, cholera, sleeping sickness all have to do with stagnant water, or presence of water bodies or improper usage of water sources. These diseases also tend to be linked to climatic factors for example in the rainy season it

is probable that proliferation of diseases like cholera, dysentery will result. Some diseases for example malaria, sleeping sickness and amoebic dysentery are both caused by protozoan parasites that enter the body through bites from flies and mosquitoes. Typhoid also presents as a disease with a lot of associations probably because it is caused by ingestion of contaminated food or water, and most other diseases have to do with a contaminated water source of some kind. Clinically typhoid especially in countries like Uganda is only diagnosed after testing for malaria and dysentery because they initially present similar symptoms of fever, etc.

```
Afpn : Maln Typn Chon
Rabn : Dysn Maln Plgn Typn Slsn Mssn Chon Gwmn Nntn
Dysn : Rabn Maln Typn Slsn Chon Plgn
Gwmn : Rabn Maln
Maln : Afpn Rabn Dysn Typn Chon Mssn Slsn Gwmn Menn
Mssn : Typn Rabn Maln Plgn Nntn
Menn : Maln
Nntn : Typn Rabn Slsn Mssn Maln
Plgn : Rabn Typn Mssn Dysn Slsn Chon
Typn : Afpn Rabn Dysn Maln Mssn Nntn Plgn Slsn Chon
Slsn : Dysn Maln Typn Rabn Plgn Nntn
Chon : Rabn Dysn Maln Afpn Typn Plgn
```

Fig. 2: Markov Blankets corresponding to graph in Figure 1 with $\alpha = 0.05$. Each row represents the target disease in the leftmost column and the corresponding Markov Blanket on the right after the semicolon.

One strange finding is the relation between rabies and the other diseases. As evident from the Figure 1 and Figure 2 there seems to be close associations between rabies and several other diseases especially malaria. From the literature there seems to be a relationship between the clinical manifestation of rabies encephalitis and cerebral malaria [Mallewa et al. 2006]. Without much certainty we conjecture that the relationship between malaria and rabies is the cause of the large Markov blanket for rabies.

Figure 4 depicts the graph obtained by decreasing the significance level to 10−4, thereby eliminating more false positives in the second phase of the IAMB algorithm as previously highlighted. The implications as seen from the graph are that there are relatively less linkages amongst the diseases. The linkages in the graph represent associations with a 10−4 chance of false positives (under the assumptions of the Z-test).

Figure 3 shows the corresponding Markov Blankets for the diseases. With a higher threshold on determining a false positive from a true positive $\alpha = 10-4$, no diseases are associated with guinea worm and meningitis. Again diseases with the highest associations present as malaria, rabies and typhoid. As previously proposed, we think malaria and rabies could be related by some external latent variable. Typhoid is related to a lot of other diseases because of similar causative environmental factors based on sanitation.

Fig. 3: Markov Blankets corresponding to graph in Figure 4 with $\alpha = 10^{-4}$. Each row represents the target disease in the leftmost column and the corresponding Markov Blanket on the right after the semicolon.

For Fisher's test with $\alpha = 0$ we obtain empty sets for the Markov blankets. With a sufficiently small $\alpha = 10-6$, we obtain the graph in Figure 6 with the corresponding Markov Blankets shown in Figure 5. We get stricter and more intuitively plausible relationships. It is interesting to note that the associative nature of rabies decreases with malaria while the associative nature of typhoid remains even with this small α. This is probably because there are stronger relationships associated with typhoid and the corresponding diseases than with rabies and malaria. This again seems to be intuitively right.

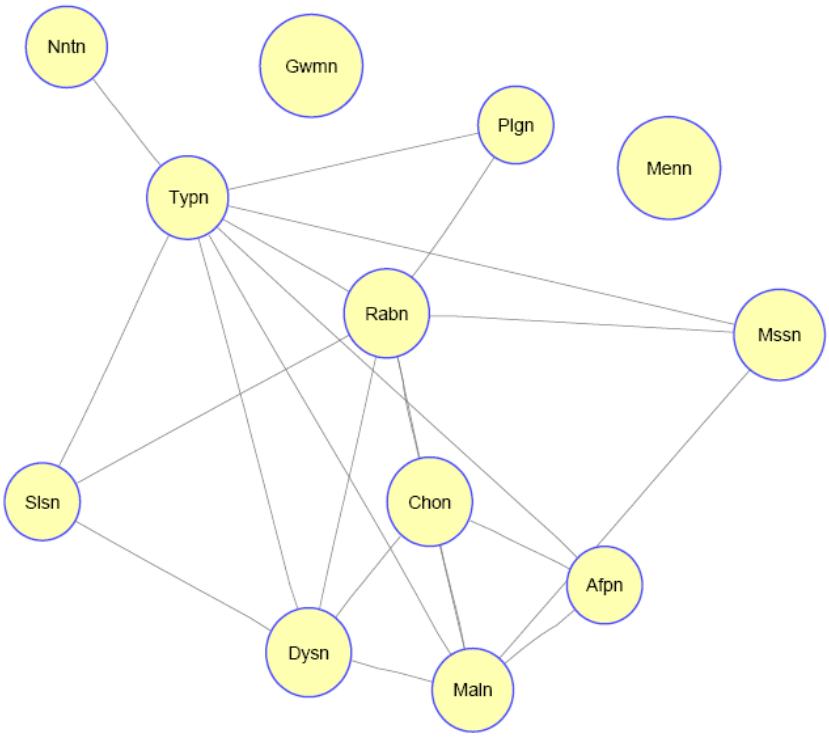

Fig. 4: Association graph showing the associative linkages between the notifiable diseases represented as : Dysn -Dysentery, Typn -Typhoid, Gwmn -Guinea Worm, Nntn -Neonatal Tetanus, Menn -Meningococcal Meningitis, Plgn -Plague, Mssn -Measles, Afpn -Acute Flaccid Paralysis, Chon -Cholera, Slsn -Sleeping Sickness, Rabn -Rabies and Maln -Malaria. $\alpha = 10^{-4}$.

5. CONCLUSION

In this paper we have shown how associations between diseases specifically notifiable diseases can be found by applying techniques from causal machine learning. From a dataset spanning over 6 years, associations amongst 12 notifiable diseases in Uganda were established. The epidemiological department at the Ministry of Health in Uganda has been making most of these associations intuitively and using experience from the medical field. In this paper, we were able to show that these associations can actually be discovered from the data; having indicated that such predictive relationships exist, these can now be exploited by being incorporated into existing disease monitoring systems.

While some of the discovered relationships seem intuitive, others are not immediately plausible and need expert domain knowledge to fully understand the relationships. This study considered all 80 districts of Uganda together pooled as one dataset. Further research will try to see if these relationships hold for different districts or different parts of the country that may have different climatic conditions. Associations will also be investigated for specific periods of the year to see if associations are more pronounced during some phases of the year than others. This study however suffices to present information for planning and preventive purposes and offers a means of determining whether predictive relationships exist between the counts of different diseases.

6. ACKNOWLEDGMENTS

We thank the Ugandan Ministry of Health for providing the disease surveillance data. The project was supported in part by the NUFFIC NPT project.

7. REFERENCES

GUYON, I., ALIFERIS, C., AND ELISSEEFF, A. 2007. Causal feature selection. In *Computational Methods of Feature Selection*, H.Liu and H.Motoda, Eds. Data Mining and Knowledge Discovery. Chapman and Hall/CRC Press, Boca Raton, FL.

MALLEWA, M., FOOKS, A., BANDA, D., CHIKUNGWA, P., MANKHAMBO, L., MOLYNEUX, E., MOLYNEUX, M., AND SOLOMON, T. 2006. Rabies encephalitis presenting as cerebral malaria. *Journal of Clinical Virology 36,* Supplement 3, S42 – S42. Programme and Abstracts of the European Society for Clinical Virology 9th Annual Meeting.

MARGARITIS, D. AND THRUN, S. 1999. Bayesian network induction via local neighborhoods. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 505–511.

PEARL, J. 2000. *Causality : Models, Reasoning and Inference*. Cambridge University Press, Cambridge.

TSAMARDINOS, I. AND ALIFERIS, C. F. 2003. Towards principled feature selection: Relevancy, filters and wrappers. In *Proc. of the Ninth International Workshop on Artificial Intelligence and Statistics.* Morgan Kaufmann Publishers.

TSAMARDINOS, I., ALIFERIS, C. F., AND STATNIKOV, E. 2003. Algorithms for large scale markov blanket discovery. In *Proc. of the 16th International FLAIRS Conference.* AAAI Press, 376– 380.

YARAMAKALA, S. AND MARGARITIS, D. 2005. Speculative Markov Blanket discovery for optimal feature selection. In *Proc. of the Fifth IEEE International Conference on Data Mining (ICDM '05).* 809–812.